Maxentropic Markov Chains

JØRN JUSTESEN AND TOM HØHOLDT

Abstract—**The Markov chain that has maximum entropy for given first and second moments is determined. The solution provides a discrete analog to the continuous Gauss–Markov process.**

I. Introduction

It is well known that random walks and similar Markov chains [1] can be used as approximations for or analogs to Brownian motion or diffusion which are more precisely described by continuous stochastic processes. In connection with digital system analysis it is also desirable to have Markov chains, which approximate continuous processes with finite variance [2], [3]. Moreover, it is an important result in information theory [4] that Gauss–Markov processes have maximum entropy among all processes with given values of the first and second moments. Another application of maxentropic Markov chains appears in [5]. In [6] the maxentropic Markov chain was used in the derivation of an upper bound to the entropy of certain spectrum shaping codes.

II. Markov Chains with Finite Variance

This paper considers Markov chains with a finite or denumerable set of states

$$S = \{\cdots, s_{-1}, s_0, s_1, \cdots\},$$

and with transition probabilities

$$P[s_k|s_j] = \begin{cases} p_{jk}, & \text{if } (j,k) \in T, \\ 0, & \text{elsewhere.} \end{cases} \tag{2.1}$$

The chain can be interpreted as representing a signal that changes as a function of time. When the chain is in state k, the signal is $x(t) = k$. The set T determines the possible transitions of the chain. This set may typically include all pairs such that $|j-k| = 1$ or all pairs for which $|j-k| \le J$ for some integer J.

We shall assume that the chain has a stationary distribution $P_k = P[s_k]$, i.e.,

$$\sum_k P_k = 1, \tag{2.2}$$

$$\sum_j P_j p_{jk} = P_k. \tag{2.3}$$

Clearly if P_k is a stationary distribution, then

$$\sum_j P_j \sum_k p_{jk}(k-j) = 0. \tag{2.4}$$

In particular we will be interested in the first and second moments of the chain, i.e.,

$$\mu = \sum_k k P_k, \tag{2.5}$$

$$\sigma^2 = \sum_k k^2 P_k - \mu^2, \tag{2.6}$$

and

$$\rho^2 = \sum_j P_j \sum_k p_{jk}(k-j)^2. \tag{2.7}$$

Manuscript received November 16, 1982; revised November 3, 1983.

J. Justesen is with the Institute of Circuit Theory and Telecommunication, Bld. 343, The Technical University of Denmark, DK-2800 Lyngby, Denmark.
T. Høholdt is with the Institute of Mathematics, Bld. 303, The Technical University of Denmark, DK-2800 Lyngby, Denmark

Here, in addition to the variance of the chain, we have specified the variance of the increment. Alternatively, we could let the correlation between successive amplitudes be given, since we have

$$\begin{aligned}\rho^2 &= E\left[(x(t+1) - x(t))^2\right] \\ &= 2(\sigma^2 + \mu^2) - 2E[x(t+1)x(t)]\end{aligned}$$

and therefore

$$E[x(t+1)x(t)] = \sigma^2 + \mu^2 - \tfrac{1}{2}\rho^2. \tag{2.8}$$

The entropy of the chain is

$$H = -\sum_j P_j \sum_k p_{jk} \log p_{jk}. \tag{2.9}$$

We shall determine the transition probabilities that maximize the entropy, subject to the constraints (2.5)–(2.7).

In [6] it was shown that the transition probabilities for the maxentropic chain in the binary case satisfy a difference equation of the form

$$p_{k,k+1} \cdot p_{k+1,k} = \alpha_k.$$

This equation can be converted to a linear second-order difference equation by using the substitution $p_k = t_k / t_{k-1}$, where $p_{k,k+1} = p_k$ and $p_{k+1,k} = 1 - p_{k+1}$. This gives the equation

$$t_{k+1} - t_k + \alpha_k t_{k-1} = 0,$$

which can be solved easily. If we express this system in matrix form we get

$$(\boldsymbol{S} - \boldsymbol{E})\boldsymbol{t} = \boldsymbol{0}.$$

In the following section we give a more general solution of a problem of this form.

III. Markov Chains with a Weight Constraint

We consider Markov chains with a given set of states $S = (s_1, s_2, \cdots, s_J)$. A transition from s_j to s_k can occur if (j,k) belongs to a finite set T. It is convenient to determine the maximum of H with respect to the probabilities $q_{jk} = P_j p_{jk}$:

$$H = -\sum_{(j,k)\in T} q_{jk} \log\left(q_{jk} \Big/ \sum_i q_{ji}\right). \tag{3.1}$$

We introduce a weight constraint of the form

$$\sum_{(j,k)\in T} P_j p_{jk} w_{jk} = \sum_{(j,k)\in T} q_{jk} w_{jk} \le W. \tag{3.2}$$

The q_{jk} satisfy the stationarity conditions

$$\sum_i q_{ji} = \sum_i q_{ij}, \tag{3.3}$$

and the consistency condition

$$\sum_{j,k} q_{jk} = 1. \tag{3.4}$$

The maximization problem is a natural generalization of a result of Shannon [7]. However, many of the details of the derivation are very similar to the calculation of the rate-distortion function for discrete memoryless sources [4]. In fact the problem stated above may be interpreted as a rate-distortion problem for a memoryless source, with the condition that the distributions of the source and reproducing alphabets are identical.

We note first that $-H$ is a convex function of the q_{jk}, so the problem can be treated as an ordinary convex program [8]. When the w_{jk} are given, there is a minimum value $W_{\min}$ of the left-hand side of (3.2), subject to the constraints (3.3) and (3.4), which may be determined by linear programming. For a certain value of W,

0018-9448/84/0700-0665$01.00 ©1984 IEEE

$W_{\max}$ the solution to the problem of maximizing H, becomes identical to the maxentropic chain without the constraint (3.2). Between these two values, H is an increasing function of W, and the solution always lies on the boundary given by equality in (3.2). It turns out that the constraints $q_{jk} \geq 0$ are always satisfied, and thus they need not be included in the variational problem.

It follows from the convexity of $-H$, that any local maximum of H is global, and we shall see that with the exception of the case $W = W_{\min}$, such a local maximum exists. We determine the maximum by introducing Lagrange multipliers, (s, ξ_k, η), where $s \leq 0$, and finding a stationary point of

$$I = -\sum_{(j,k)\in T} q_{jk} \log\left(q_{jk}/\sum_i q_{ji}\right) + s\sum_{(j,k)\in T} q_{jk}w_{jk} + \xi_k\left(\sum_i q_{ik} - \sum_i q_{ki}\right) + \eta\sum_{(j,k)\in T} q_{jk}. \quad (3.5)$$

We get

$$\frac{\partial I}{\partial q_{jk}} = -\log p_{jk} + sw_{jk} + \xi_j - \xi_k + \eta,$$

and thus

$$p_{jk} = \exp(sw_{jk})v_k/\lambda v_j, \quad (3.6)$$

where

$$v_k = \exp \xi_k \quad \text{and} \quad \lambda = \exp(-\eta).$$

When these transition probabilities are used, the relations

$$\sum_k p_{jk} = 1, \qquad \sum_j P_j p_{jk} = P_k$$

become

$$\sum_k \exp(sw_{jk})v_k = \lambda v_j, \quad (3.7)$$

and

$$\sum_j P_j \exp(sw_{jk})/v_k = \lambda P_k/v_k. \quad (3.8)$$

We may interprete (3.7) and (3.8) as indicating that the vectors $\{v_j\}$ and $\{P_j/v_j\}$ are right and left eigenvectors of the matrix $Q = \{\exp(sw_{jk})\}$ and λ is the corresponding eigenvalue.

By applying the well-known argument [4] we see that $H + sW$ is maximized and that this quantity is the H-axis intercept of the line with slope $-s$ through (W, H). Thus $-s$ is the slope of the tangent to the H-versus-W curve in this point. Using (3.6) we can calculate H as

$$\begin{aligned} H &= -\sum_j P_j \sum_k p_{jk}(sw_{jk} + \log v_k - \log v_j - \log\lambda) \\ &= \log\lambda - sW + \sum_j\sum_k P_j p_{jk}(\log v_k - \log v_j) \end{aligned}$$

so that

$$H = \log\lambda - sW. \quad (3.9)$$

Thus $\log\lambda = H + sW$ is maximized, and we conclude that λ is the largest eigenvalue of the nonnegative irreducible matrix Q. This also ensures, at least in the finite case, that the P_j's and the vector $\{v_j\}$ are positive. We can now summerize the results as follows.

Theorem 1: Let the set of states and the set of possible transitions of a Markov chain be given, and let a set of weights w_{jk} be associated with the transitions. The maximal entropy H of the chain is then determined parametrically as a function of the average weight W by

$$H = \log\lambda - sW,$$

where λ is the largest eigenvalue of the matrix $Q = \{\exp(sw_{jk})\}$. The transition probabilities of the maxentropic chain are

$$p_{jk} = \exp(sw_{jk})v_k/\lambda v_j,$$

where $\{v_j\}$ is the right eigenvector corresponding to the eigenvalue λ, and the stationary probabilities are

$$P_j = v_j v_j^*,$$

where $\{v_j^*\}$ is the corresponding left eigenvector.

We note that the function $H(W)$ is enveloped by the family of straight lines passing through the point $(0, \lambda)$ with slope $-s$ as s runs through the interval $[-\infty, 0]$. For all values of $s > -\infty$ the eigenvector $\{v_j\}$ is strictly positive, by the Perron–Frobenius theorem, and thus $p_{jk} > 0$. Thus a chain with a reduced number of states can occur only for $s = -\infty$, corresponding to $W = W_{\min}$.

We shall now apply the above results to the specific problem discussed in Section II. We prefer to rewrite the constraints on the moments as

$$\frac{1}{2}\sum_{j,k} q_{jk}(j+k) = \mu, \quad (3.10)$$

$$\sum_{j,k} q_{jk}(j-k)^2 \leq \rho^2, \quad (3.11)$$

and

$$\frac{1}{2}\sum_{j,k} q_{jk}(j^2+k^2) \leq \sigma^2 + \mu^2 = \sigma'^2. \quad (3.12)$$

The problem is simplified if we consider H to be a function of μ, ρ^2 and σ'^2 rather than σ^2. Theorem 1 can be extended to include several constraints, and we find that the maximal entropy is obtained by taking

$$p_{jk} = Q_{jk}v_k/\lambda v_j, \quad (3.13)$$

where $\{v_j\}$ is the eigenvector corresponding to the maximal eigenvalue λ of the matrix

$$Q = \{Q_{jk}\} = \begin{cases} \exp\left(\frac{1}{2}s(j^2+k^2) + \frac{1}{2}r(j+k) + t(j-k)^2\right), \\ \quad (j,k) \in T, \\ 0, \quad \text{elsewhere}, \end{cases} \quad (3.14)$$

where $s \leq 0$, $t \leq 0$. The entropy is

$$H = \log\lambda - s\sigma'^2 - r\mu - t\rho^2 \quad (3.15)$$

and $-(s, r, t)$ are partial derivativies of H with respect to σ'^2, μ, and ρ^2.

It follows from the Perron–Frobenius theorem that the maximal eigenvalue of the Q matrix is unique, and that the corresponding eigenvector has positive coordinates, so for the maxentropic chain all states and all transitions have positive probabilities. Consequently, if more states are added, we get a strictly larger entropy. It can be proved that for the infinite Q matrix a Perron–Frobenius theorem still holds [9], [10], and that the chain defined by (3.13) with an infinite number of states is the maxentropic chain.

References

[1] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1. New York: Wiley, 1950, ch. 14.
[2] D. Slepian, "On maxentropic discrete stationary processes," *Bell Syst. Tech. J.*, vol. 51, pp. 629–653, Mar. 1972.
[3] Y. Linde and R. M. Gray, "A fake process approach to data compression," *IEEE Trans. Commun.*, vol. COM-24, pp. 840–847, June 1978.

[4] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice Hall, 1971.

[5] D. Kazakos, "Robust noiseless source coding through a game theoretic approach," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 576–583, July 1983.

[6] J. Justesen, "Information rates and power spectra of digital codes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 457–473, May 1982.

[7] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, 1948.

[8] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton University, 1970, sect. 28.

[9] E. Seneta, *Nonnegative Matrices*. London: George Allan and Unwin, 1973.

[10] J. Justesen and T. Høholdt, "Maxentropic Markov chains," rep., Institute of Mathematics, Technical University of Denmark, 1983.

An Achievable Region and Outer Bound for the Gaussian Broadcast Channel with Feedback

LAWRENCE H. OZAROW, MEMBER, IEEE, AND S. K. LEUNG-YAN-CHEONG

***Abstract*—A deterministic coding scheme is presented for additive white Gaussian broadcast channels with two receivers and feedback from the receivers to the transmitter. The region of data rates at which reliable communication is possible is larger than that of the corresponding channel without feedback. This is the first example showing that the capacity region of degraded broadcast channels can be enlarged with feedback.**

I. Introduction

It is well known that the capacity region of discrete-time memoryless multiple access channels can be enlarged by the use of feedback [1]–[3]. This effect is intuitively reasonable, since the introduction of feedback completes a pair of communication links between the transmitters, allowing collaboration.

Until the appearance of [5] a similar enlargement of broadcast channels had not been reported. Indeed, since only one transmitter is present in the broadcast situation, it had not been at all clear that enlargement should occur. Furthermore, El-Gamal [4] had shown that for cascaded broadcast channels ("physically degraded" channels) it does not.

In [5] Dueck showed by example that the capacity region of broadcast channels can be enlarged with feedback. In this correspondence we will give a constructive scheme for additive Gaussian broadcast channels which, except in the physically degraded case, enlarges the capacity region over that obtained without feedback. Since all Gaussian channels are degraded (generalizing the usual definition of degraded channels to encompass continuous alphabets), these results show that the negative result obtained in [4] does not extend beyond the physically degraded case.

The coding scheme described here is similar to one presented in [3] for the Gaussian multiple access channel, and both are based on the scheme proposed by Schalkwijk and Kailath [6], [7] for the band-limited Gaussian channel with one transmitter and one receiver. While the scheme in [3] was shown to achieve the capacity region of the multiple access channel, no such result is claimed for the broadcast channel. Although our achievable region will be shown to lie surprisingly close to an easily obtained (and apparently optimistic) outer bound for some cases, determining the actual capacity region of the Gaussian broadcast channel with feedback remains an open problem.

Manuscript received January 4, 1981.

L. H. Ozarow is with AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974.

S. K. Leung-Yan-Cheong is with the University of British Columbia, Vancouver, BC, Canada.

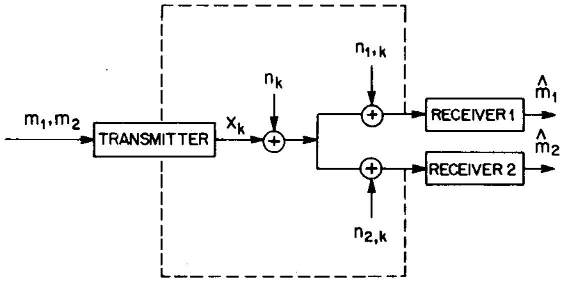

Fig. 1. Gaussian broadcast channel with feedback.

II. Gaussian Broadcast Channel

Consider the channel shown in Fig. 1. A single transmitter wishes to send separate data to two receivers using the discrete-time additive white Gaussian channel shown. At any time the transmitter has available the prior channel outputs of both receivers. We assume that information is transmitted on a block basis, and that for each block we impose an average power constraint of the form

$$\frac{1}{N}\sum_{k=1}^{N}\overline{x_k^2} \le P,$$

where the expectation (denoted by an overbar) is carried out over the ensemble of messages and channel noise. If m_1 and m_2 are the messages to be sent to the first and second receiver respectively, then the rates achieved are defined by

$$R_i = \frac{1}{N}\log M_i \text{ nats/transmission}, \qquad i = 1,2,$$

where $M_i = \|m_i\|$, that is the number of possible values of m_i. We take all logarithms to be natural, so that rates are given in natural units.

A rate pair (R_1, R_2) is achievable if for any $\epsilon > 0$, there exists a code with some sufficiently large block length N, for which the number of codewords to each receiver satisfies

$$M_i \ge e^{NR_i}, \qquad i = 1,2,$$

and

$$\Pr[\text{receiver } i \text{ incorrectly decodes } m_i \text{ for } i = 1 \text{ or } 2] < \epsilon.$$

The capacity region of the channel is the closure of the set of achievable rate pairs.

Returning to the channel, as shown in the figure, the real number x_k is corrupted by the addition of a common noise component n_k and two separate noise components $n_{i,k}$ $(i = 1,2)$ with respective variances σ^2, σ_1^2, and σ_2^2. We assume that each noise process is white and that the three are mutually independent. The capacity region of the channel in the absence of feedback is the subset of the positive quadrant bounded by the R_1 and R_2 axes, and the curve are described parametrically by

$$R_1 = \frac{1}{2}\log\left(1 + \frac{\alpha P}{\sigma^2 + \sigma_1^2}\right),$$

$$R_2 = \frac{1}{2}\log\left(1 + \frac{(1-\alpha)P}{\sigma^2 + \sigma_2^2 + \alpha P}\right), \tag{1}$$

where $0 \le \alpha \le 1$. This characterization was obtained in [8] and [9].

0018-9448/84/0700-0667$01.00 ©1984 IEEE